Data Pre-processing for Database Marketing

Filipe Pinto[1], Manuel Filipe Santos[2], Paulo Cortez[2], and Hélder Quintela[2]

[1] Department of Computer Science Engineering, School of Business and Technology, IP Leiria, Portugal
fpinto@estg.ipleiria.pt
[2] Department of Information Systems, University of Minho, Portugal
{mfs, pcortez, hquintela}@dsi.uminho.pt

Abstract. To increase effectiveness in their marketing and Customer Relationship Manager activities, many organizations are adopting strategies of Database Marketing (DBM). Nowadays, DBM faces new challenges in business knowledge since current strategies are mainly approached by classical statistical inference, which may fail when complex, multi-dimensional and incomplete data is available. An alternative is to use Knowledge Discovery from Databases (KDD), which aims at automatic extraction of useful patterns by using Data Mining (DM) techniques. When applied to DBM, the identified patterns can be used for the efficient characterization of the customers. This paper focus several problems that arose in the data pre-processing step (e.g. data cleaning), which is necessary for the success of the DM approach to a DBM project.

1 Introduction

The concepts of mass producing and mass marketing, created during the Industrial Revolution and exploited up to nowadays, is being challenged by the new approach of one-to-one marketing. The Database Marketing (DBM) activity has changed significantly over the last years. In past, database marketers applied business rules to target customers directly, based sometimes in the marketer's intuition. The current approach relies on predictive response models to target customers for offers. These models accurately estimate the probability that a customer will respond to a specific offer and can significantly increase the response rate to a product offering. The old model of *design-build-sell* is being replaced by *sell-build-redesign*.

Due to advances in information and communication technologies, corporations can effectively obtain and store transactional and demographic data on individual customers at reasonable costs. The challenge now is how to extract important knowledge from these vast databases. Through a process called Knowledge Discovery from Databases (KDD), organizations can empower the stored data, understanding the customers' preferences and behaviours through analyzing their transactional data.

However, in almost cases, the data set presents several problems as the result of procedural factors, inadequate questionnaire options, refusal of response, or/and inadequate database schemas. The most part of DM methodologies (e.g., CRISP-DM, SEMMA) deal with these data quality problems in their initial phases (e.g., data

understanding, data selection, data cleaning and data transformation). Several studies emphasis that almost 80% of KDD effort is spent in data related phases. Some marketing studies outlines that data quality problems cost 10% of the total revenue [1].

This paper presents a case study of a DBM project carried out by a Portuguese marketing enterprise, evidencing the problems surged in the data pre-processing process. The referred company distributes an own-branded magazine which includes discount vouchers to promote products of a great multinational distribution organisation (food and beauty products).
The main goal of the project was to complete the cycle from the voucher utilisation in the supermarkets to the product and customer association. Such association rules can be used as filters on the databases in order to identify the candidates to buy determinate products.
The further sections are organized as follows: first, some concerns about data quality for DBM are presented; then, a case study of pre-processing data in a DBM project that involved several pre-processing tasks (e.g., merging of two DB, reducing/eliminating inconsistencies, levelling attributes) is described; finally, closing conclusions are drawn.

2 Database Marketing

2.1 Basic concepts

Customer Relationship Management (CRM) is defined by four elements of a simple framework: *Know*, *Target*, *Sell* and *Service* [2]. It enables organizations to know and to understand its markets and customers. This involves detailed customer intelligence in order to select the most profitable customers and identify those no longer worth targeting. CRM also entails development of the offer: which products to sell to which customers and through which channel. In selling, firms use campaign management to increase the marketing department's effectiveness. Finally, CRM seeks to retain its customers through services such as call centers and help desks.
CRM is a combination of several components. Before the process can begin, the organization must possess customer information. Companies can learn about their customers through internal customer data or they can purchase data from outside sources (e.g., billing records, customer surveys, web logs, credit card records). For that, company data warehouses are a critical component of a successful CRM strategy [3]. Most companies have vast databases, with poor data quality.
From a marketing perspective, the CRM activity can be viewed as a process, known as DBM, to establish profitable interaction with the clients. Currently DBM is mainly approached by a classical statistical inference, which may fail when complex, multi-dimensional, and incomplete data is available.
In DBM there are two critical components: customer data and customer KDD. The customer data can be obtained by different interactions (e.g., surveys, voucher discounts), by transactions and/or by rent of external databases. Great amounts of data

are important for the generation of accurate customers' patterns (e.g., preferences, behaviour, segmentation) since measures as relevance and quality are taking in account. Customer KDD plays a critical role in the overall DBM process because it implies interaction with the data mart or data warehouse in one direction, and interaction with marketing function in the other direction. Another important issue is the link between KDD applications and campaign management software, which should be automatic to ensure data and models integrity.

Activities of DBM can be grouped in three different stages: *Data Warehousing*, *KDD*, and *campaign execution*. *Data warehousing* is the storage of information about customers such as: purchase history, product returns, phone and mail contacts, click-stream patters, credit and payments. *KDD* aims the automatic pattern extraction using DM techniques. The KDD process involves several phases (e.g., Business Understanding, Data Selection, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment and Monitoring). The dataset plays a central role in the KDD process and the success of KDD heavily depends of the implementation of good strategies to guarantee the quality of data. *Campaign execution* is the final stage when the marketer implements the recommendation action, (e.g., mails personalized offers to selected customers), collects feedback of the responses and updates the transaction database.

2.2 Data Quality

The success of a DBM strategy depends highly of data warehouse availability and quality. Data quality is a multidimensional concept [4] as data itself is multidimensional [5][6]. Modern definitions of data quality have a wider frame of reference and many more attributes than the obvious characteristics of accuracy. One of the most interesting views about data quality, widely adopted by literature, takes a consumer point of view: "data that is fit for use" [7], [8], [9], [10] or in other words: "Data are of high quality if they are fit for their intended uses in operations, decision-making, and planning. Data are fit for use if they are free of defects and possess desired features" [11].

Another perspective about data quality notes that data considered appropriate for one use may not possess sufficient quality for another use (e.g. the case of multiples uses of data through data warehouses). Therefore, a data quality strategy in a DBM program must consider the end user and allow that user to define the appropriate level of data completeness and cleanliness. Requirements may be different for corporate data than for local data. First steps in any improvement process must be to identify the uses made of the data and by whom. A data quality strategy also needs to look forward to the future potential uses of the data.

The problem of missing or inconsistent data has been a pervasive problem in data analysis since the origin of data collection [12]. Missing data and data quality regarding data warehousing and CRM it is an area of recent research. It should be stressed that the use of noise, irrelevant and missing data in the DM phase conduces to the identification of inadequate and irrelevant patterns; and in the majority of

situations it increases the difficulty in the learning phase. The impact of missing data depends on which DM algorithm is being used.

3 Pre-processing in DBM: A case study

DBM is characterized by enormous amounts of data at the level of the individual consumer. However, these data have to be turned into information in order to be useful. In this section we present the work developed in order to get a DB with high data quality by means of data clean and data pre-processing tasks.
The aim of the case study presented in this paper is to segment customers, using a clustering approach, with the intention of finding a set of simple rules which explain clusters of clients with homogeneous behaviours. For this study, the dataset was created after a direct marketing project (detailed in following sections) where discount vouchers have been offered to thousands of potential clients, through a own branded magazine. A specific hygienic product was targeted, since it presented high sales.
In this paper, a strong emphasis is dedicated to the validation and elimination of irrelevant data and extensive data pre-processing, corresponding to the pre-processing phase of the KDD project. After a brief explanation about the marketing data available, some concepts will be introduced about clean data and then each of its steps will be described.

3.1 Marketing Data

In this case-study two DB of the same organization were used: one containing personal information about registered customers and another with the transactional data with the voucher used at the supermarkets (Figure 1). The first database was created by renting an external DB and merging it with others from previous direct marketing projects. The imported data from external DB includes only the customer's name and address.

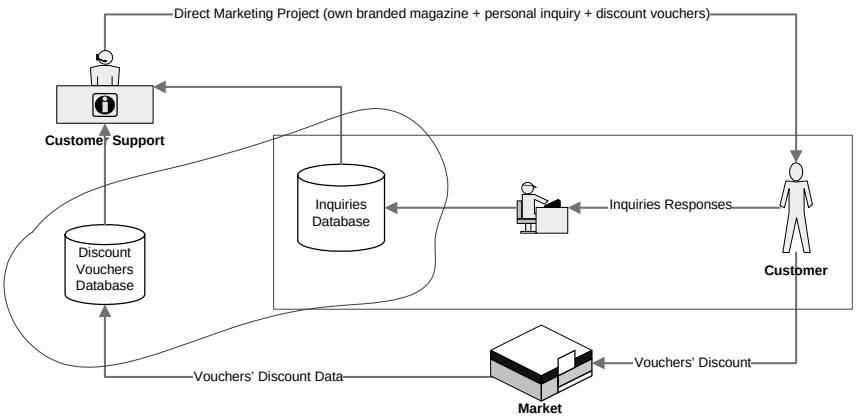


Fig. 1. Data Acquisition Schema

The first contact takes place via one own-branded magazine which includes discount vouchers (Figure 1). These vouchers are from different kinds of products and different values. This contact with the customer aims at establishing a direct contact (by postal address) with them, allowing the organization to receive a fulfilled questionnaire which conveys some individual information, composed of five items described in Table 1.

Only after the voucher use is possible to close the circuit and know which products fits customer needs and simultaneously customers' profile. The second database contains transactional data with registered discount of the vouchers in the supermarkets. This project focused on the Most Valuable Consumers, which are classified by taking into account the domestic appliances (e.g. household, dishwater and washing machine).

Table 1. Questionnaire main attributes

Attribute	Domain
1. Household	{Non response, 1, 2, 3, 4, 5, 6 or more}
2. Dishwater	{Non response, Yes, No}
3. Monthly Consumption (€)	{Non response, [0…150], [151…350], [351…500], [501…650], [651…[}
4. Household Income (€)	{Non response, [0…500], [501..750], [751…1000], [1001…1500], [1501…2250], [2251…[}
5. Childs	{Non response, No, Yes}
6. Number of Childs	{Non response, 1, 2, 3, 4, 5, 6, 7, 8, 9, [10…[}

3.2 Cleaning work

The first database handled during this project was the one that keeps personal information about each contact (prospect). This DB had initially more than 600 thousands registers (name and address) and the first step was to know which of them were of interest for the project. The organization decided that the only data to be included in the study should be that of the costumers that responded to the questionnaire. The number of registers to be considered came down to 253 thousand.

As referred above there's another DB with transactional data. This DB registers all transactions made by the customer, i.e., all vouchers discounted and all answers to all questionnaires sent. Using the information in this database we found out those costumers that had used some vouchers and only those were selected from the contacts database resulting in a database with 64 482 contacts.

Both databases of this project presented several problems, mainly in terms of the data questionnaires. The transactional DB contained several missing data, which had to be treated. However, this drawback did not occur in the discount vouchers DB, due to the automatic acquisition of the data. After the merging process, a single dataset was created, and the DBM project developed.
The registers that presented some inconsistency in attribute values were deleted. The most relevant were *age* and *sex* with outliers. Every record containing the some problems was rejected resulting in a database with 63961 "clean contacts".

3.4 Pre-Processing work

To solve the problem of missing data, some techniques were used, taking in account several aspects regarding statistical significance issues after pre-processing the data.

3.4.1 Missing values

The first six attributes in several registers contained *blank values*, being a possible cause the incorrect recording of the costumer's response. In the Table 1 is stated that one of the possible values for the client's response was (*Non response*). For some reason this non response was registered using blank values for the customers that did not responded to the addressed question. This is due to a refusal when some respondents find some questions personally or sensitive (e.g., political, religious affiliation, education level, income, age) and procedural factors (human factor) in the introduction in DB.
The *blank values* were considered *Non Response*, assuming an error in introduction of data by the operator, reinforcing the necessity of validation mechanisms in the software used.

3.4.2 Data consistency and De-duplication

In this customer database were detected multiple variations of the same value like city, zip code, company, customer or address including multiple abbreviations and types. Prevalence of such type of inconsistency leads to the problem of duplication of records.

Table 2. Example of multiple variations for similar values

Country	Zip code	City
PT	2400	Vila Nova Gaia
PT	2400-230	Gaia

Considering the last row of Table 2, and assuming this address value to belong to "ESTG - Gaia", duplicate records will be added for this company for the same location (i.e., ESTG – Vila Nova de Gaia). Such duplication can lead to data integration problems. Consequently, if we try to join two tables on a particular

attribute value (as given in the current example), 'Vila Nova de Gaia' will not come out to be equal to 'Gaia', even they are same. In this case that was necessary to find manually each case evolving two or more records and correct them with one unified value.

3.4.3 Data conformance

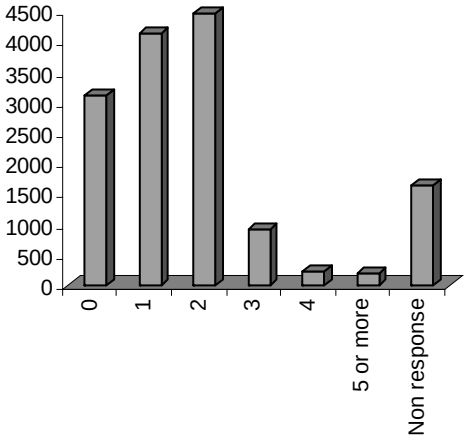
Some data conformance faults were detected between fields with values which must make sense, as *Childs* and *Number of Childs*. These Attributes were processed together to guarantee the validation of data (Table 3). For example, when the value for attribute *Childs* is equal to *N*o and *Number of Childs* equal to *Non Response* (3004 registers), the value of *Number of Childs* was changed to 0. The cases in where the value for attribute *Childs* was equal to *Yes* and *Number of Childs* was equal to *0* or *Non Response*, or where the value for attribute *Childs* was equal to *No* and *Number of Childs* not equal to *0* or *Non Response*, were discarded.

Table 3. Answers to questions *Childs* and *Number of Childs*

	Number of child's											
Childs (yes/no)	0	1	2	3	4	5	6	7	8	9	10 +	Blank
Yes	9	4213	4597	929	236	57	24	8	4	4	20	224
No	222	8	4	5	1	2	0	0	0	0	1089	3004
Non response	2	65	56	12	3	1	0	0	0	0	10	785
Blank	23	0	0	0	0	0	0	0	0	0	77	3688

In the attribute *Number of Childs* were detected too much classes. A significant bias occurs due to natural and inadequate questionnaire options. A work of levelling was done reducing the number of classes (Fig. 2).

Fig. 2. Attribute Number of Childs distribution after a leveling process



4 Conclusions

Ideally, the targeted DBM project data should be integrated in a Data Warehousing, which would facilitate the KDD process. However, this was not the case. The original data was gathered in different databases with different purposes. Therefore, this paper brought to the light some important issues that needed to be addressed in the data pre-processing stage, such as:

- Selection of data - the careful study and identification of the useful data tables;
- Data processing - the correct interpretation of the existing relations among the tables and the construction of a unique table for DM;
- Data quality – the statistical analyses of the data to find problems (e.g., wrong values)
- Data cleaning – the selection of most valuable consumers, which received and response to the inquiries;
- Missing values – the analysis and correction (e.g., imputation, delete) of attributes with blank values;
- Data consistency and De-duplication – the verification of consistency for same instances and duplicate entries of an entity;
- Data conformance – the study of related attributes to verify the expected conformance between.

Some preliminary DM experiments were performed with the original data, leading to irrelevant patterns. After a careful analysis, the pre-processing (described in this

paper) stage was endured. Then, a new DM modelling was performed, using a clustering approach, where a Self-Organizing Map was used to segment the clients. Next, a decision tree (C5.0 algorithm) was applied to each cluster, in order to generate a set of explanatory rules. Under this scheme, a classification accuracy between 75% and 82% was achieved.

Further work will be made in order to prove the correctness of the decisions taken in this project. The knowledge obtained through the DM process will be represented and archived in a knowledge based system to support a direct contact to the customers and to promote a restricted pre-selected set of products. The feedback obtained will be registered and used to corroborate or to refute some of the pre-processing operations and to form a meta-knowledge level to guide future KDD projects.

References

1. Laudon K. C., Data Quality and due Process in Large Inter-organizational Record Systems, Communications of ACM. 29, 1, 4-11, 1986.
2. IDC & Cap Gemini. Four elements of customer relationship management. Cap Gemini White Paper.
3. Yen, D.C., Wang, J.C. and Rygielski, C., Data Mining techniques for customer relationship management, Technology in Society, 24, pp. 482-502, 2002.
4. Klein, B. and D.F. Rossin, Data errors in neural network and linear regression models: An experimental comparison. Data Quality, 1999. 5(1): p. 25.
5. Laudon K. C., Data Quality and due Process in Large Inter-organizational Record Systems, Communications of ACM. 29, 1, 4-11, 1986.
6. Juran, J.M. and A.B. Godfrey, Juran's Quality Handbook. 5 ed. 1999, New York: McGraw-Hill.
7. Strong, D.M., Y.W. Lee, and R.Y. Wang, Data quality in context, Communications of the ACM, 1997. 40(5): p. 103-110.
8. Wang, R.Y., D.M. Strong, and L.M. Guarascio, Beyond Accuracy: What data quality means to data consumers. 1996, Total Data Quality Management Programme.
9. Department of Defence, U.S, DOD Guidelines on Data Quality Management. 2003, Defence Information Systems Agency. p. 28.
10. Brown, S., Data quality: Relatively critical and critically relative. DM Review, 2002.
11. Tayi, G.K. and D.P. Ballou, Examining data quality. Communications of the ACM, 1998. 41(2): p. 54-57.
12. Brow, M.L. and Kros, J.F., Data Mining and the impact of missing data, Industrial Management & Data Systems, 103/8, pp. 611-621, 2003.